ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 6, Issue 1 (Jan. - Mar. 2018)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2018)

# Applying Data Mining Techniques in Taxes Sales

[I]Sally S. Elesnawy, [II]Mohhamed A. El-Dosuky, [III]Hazem M. El-Bakry, [IV]Aziza S. Asem
[I,II,III,IV]Faculty of Computer and Information Sciences, Mansoura University, Egypt

**Abstract**
*This paper investigates how tax administrations could make use of data mining and econometrics. The operational framework is proposed. It performs time series analysis by regression. It attempts to handle inherent data problems such as missing, duplicate, outlier and chronological data.*
*First, missing entities make generalization difficult. We propose using Interpolation to handle this problem. Second, Duplicated data where an attribute has two or more identical values. So we propose normalizing the data. Third, tax data outlier may lead to wrong conclusions. We propose splitting data into sets. Forth, data are scattered into chronological format so features are not ready to match standards. We propose using feature Engineering.*

**Keywords**
*Econometrics, Data Mining, Interpolation, feature engineering, time series, regression.*

## I. Introduction

Taxes are vital for GDP. There is a dichotomy of taxes into either direct or indirect [ 2]
There are many **Problems** in taxes data, such as:

- ***Duplicated data*:** is Present when an attribute has two or more identical values. So we should normalized data [1].
- Tax Data usually has missing entities this makes generalization difficult we propose using ***Interpolation*** [9].
- Tax Data is not ready in form of features rather it is scattered into chronological so features are not ready to match standards we propose using ***feature Engineering*** [3].
- Tax data outlier analysis so we need to split data into sets. [4]

The remaining sections are: 2nd section is previous work. 3rd section is the proposed system. 4th section shows the evaluation. 5th section is conclusion.

## II. Previous Work

In a recent paper [6], an empirical proof is given to detect taxpayers with false invoices in Tax Administration of Chile using data mining techniques, such as clustering algorithms, decision trees, and Bayesian networks.
In a recent paper [10],is that there is no one single tax system that can be qualified as
the optimized tax regime in Suriname. The current mining taxation regime is relatively low risk for the government, but does not ensure maximum financial benefits in case of high commodity prices.
Basically speaking, taxes data can be either ***cross-sectional*** or ***time series data*** (TSD) [5,11,12]. TSD is a set of observations over time that requires special care to the data frequency at which the data are collected. Some data sets have both cross-sectional and time series features such as pooled cross section or panel data. Reaching to a model for cross-selection data using regression analysis is relatively simple compared to time series modeling (TSM) [5,7]. There are many computer tools that are commonly used for the TSM such as NLOGIT (www.nlogit.com ) and RATS (www.estima.com )

## III. Proposed system

**Algorithm 1: Preprocessing algorithm**

```
Preprocessing ()
        T=load tax Data ().

        If T is not normalized
                T=Normalize (T)
        End If.

        If Has Missing Data (T) Then
                T=Interpolation (T)
        End If.

        // Outlier Analysis
        Mean = get_mean()
        Range = get_range()
        T = Remove Outlier ( T, Mean, Range)
        Save T.

        Cs=clustering (T).
                For each c in Cs
                        Save C.
                End for each.
```

Algorithm 1 show that we load tax data and load all requirements but after we made Normalization to data and determine if data has missing data so use Interpolation and if data outlier analysis data so use clustering and if data has chronological data so we will use Feature Engineering

### A. Normalization

It is the transformation data into related tables to save typing of repetitive data [4]
Table1 shows Normalization Example Taxes Clients.

Table1: Arbitrary Taxes Clients

| T_id | T_Name | T_Address | Taxespayer_Type |
|---|---|---|---|
| 401 | Adam | Noida | sales |
| 402 | Ahmed | Panipat | Sales and tables |
| 403 | Shawky | Jammu | Tables |
| 404 | Adam | Noida | tables |

Normalization rule are divided into following normal forms: 1st Normal Form(1NF) as in Table3 , 2nd Normal Form (2NF)as in

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

both Table4 and Table5, and 3rd Normal Form (3NF) as in Tables from 6 to 8.

Table 2: Taxes Payer Table

| T_Name | Tax | Subject |
|---|---|---|
| Adam | 15 | Sales,tables |
| Ahmed | 14 | Sales and tables |
| Shawky | 17 | Tables |

Table 3: Taxes payer Table following 1NF.

| T_Name | Tax | Subject |
|---|---|---|
| Adam | 15 | Sales |
| Adam | 15 | Tables |
| Ahmed | 14 | Sales and Tables |
| Shawky | 17 | Tables |

Table 4: New Taxes payer Table following 2NF

| T_Name | Tax |
|---|---|
| Adam | 15 |
| Ahmed | 14 |
| Shawky | 17 |

In Tax Table the candidate key will be **T_Name** column, because all other column i.e. **Tax** is dependent on it.

Table 5: New Subject Table introduced for 2NF

| T_Name | Subject |
|---|---|
| Adam | Sales |
| Adam | Tables |
| Ahmed | Sales and Tables |
| Shawky | Tables |

Table 6: Taxes_Payer_Detail Table

| Taxes Payer_id | Taxes Payer_ name | DOB | Street | city | State | Zip |
|---|---|---|---|---|---|---|

Table 7: Taxes payer Detail Table

| Taxes Payer _id | Taxes Payer_name | DOB | Zip |
|---|---|---|---|

Table 8: Address Table

| Zip | Street | city | state |
|---|---|---|---|

### B. Interpolation

It is the construction of brand new data within a range [9]

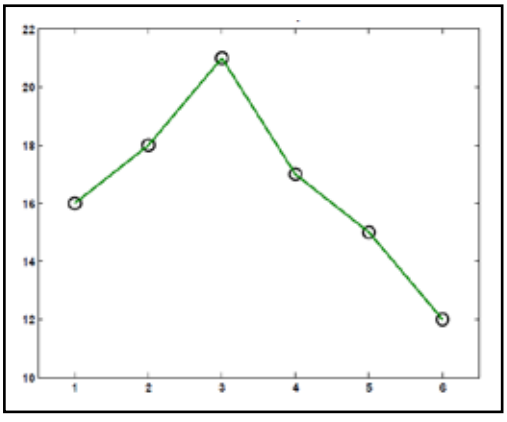

Fig.1 : Piecewise linear interpolation

### C. Feature Engineering

It is the redefinition of the problem features for more usability[3]. Transforming birth_date into age worths a honorable mention.

### D. Regression

To logarithmically devise relation from independent to dependent variables [8]. The linear model concerning $y = mx + b$ is depicted in Figure2.

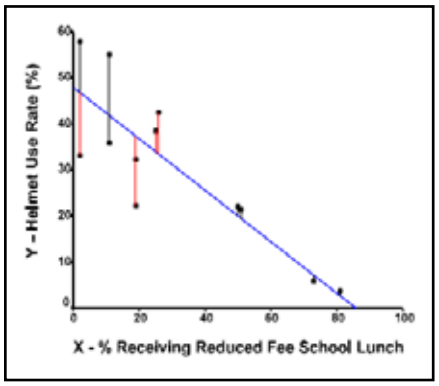


Fig. 2 : linear regression

Likelihood Estimation can be formalized using Akaike Information Criterion or R-Squared

$$L(\theta|x_1,\ldots,x_n) = f(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

## IV. Evaluation

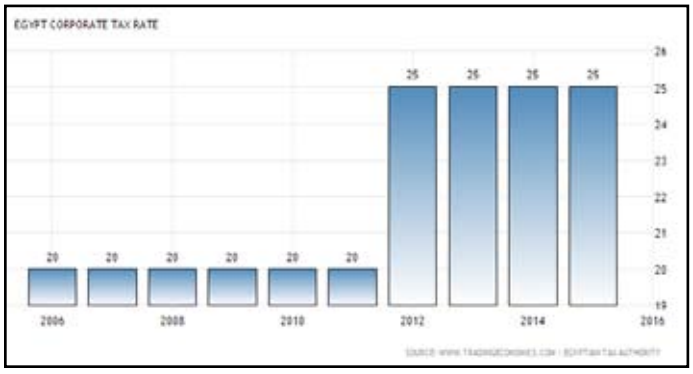


Fig. 3 : Data From OCED.org

| Location | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|---|---|
| Australia | 11.504 | 11.936 | 11.795 | 11.912 | 12.169 | 11.869 |
| Austria | 9.301 | 9.905 | 9.664 | 9.68 | 9.433 | 8.965 |
| Belgium | 13.704 | 13.995 | 13.866 | 13.61 | 13.399 | 12.618 |
| Canada | 12.851 | 12.842 | 11.524 | 11.212 | 11.244 | 11.471 |
| Czech Republic | 4.201 | 4.113 | 4.298 | 4.477 | 4.458 | 4.234 |
| Denmark | 24.915 | 25.235 | 24.929 | 24.864 | 24.245 | 24.204 |
| Estonia | 6.83 | 6.516 | 6.43 | 6.479 | 6.265 | 5.526 |
| Finland | 14.031 | 13.589 | 13.526 | 13.133 | 12.732 | 12.903 |
| France | 7.755 | 7.567 | 7.293 | 7.338 | 7.176 | 7.717 |
| Germany | 9.187 | 9.51 | 8.646 | 8.28 | 7.766 | 7.818 |
| Greece | 4.797 | 4.348 | 4.398 | 4.199 | 4.271 | 4.581 |

Fig. 4 : sample data of Tax Income

It is noticed that some country decreased the taxes on every year

Open TSM (Time Series Modeling), Then Load Data, Make Data Transformation and Editing. Should make 1-select option 2-select variables 3-Press Go

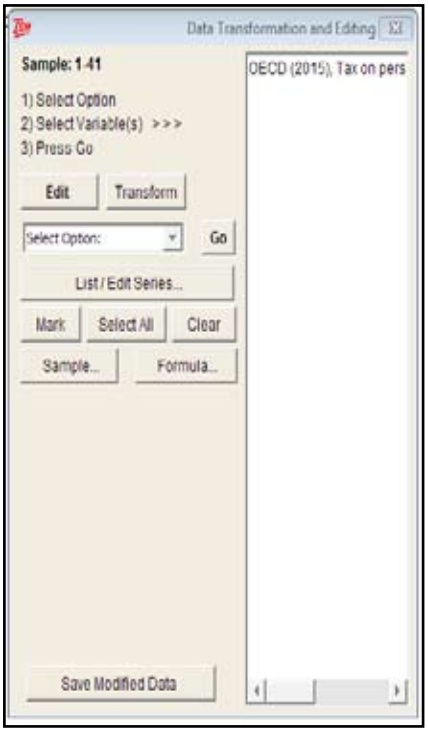


Fig. 5 : TSM settings

To plot data choose (series Plot, correlogram, partial correlogram, specturm)

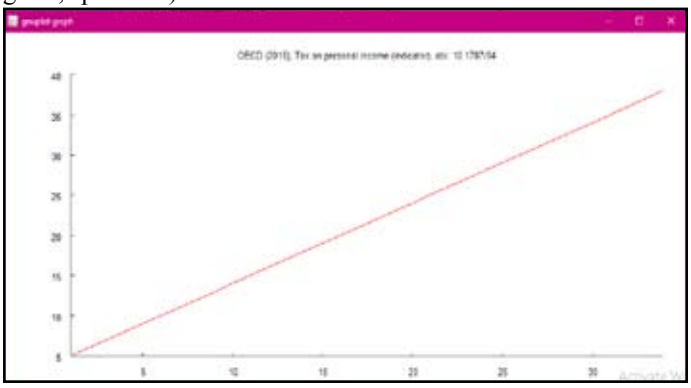

Fig. 6 : Series Plot

Figure 7 shows that increased the line degrade.

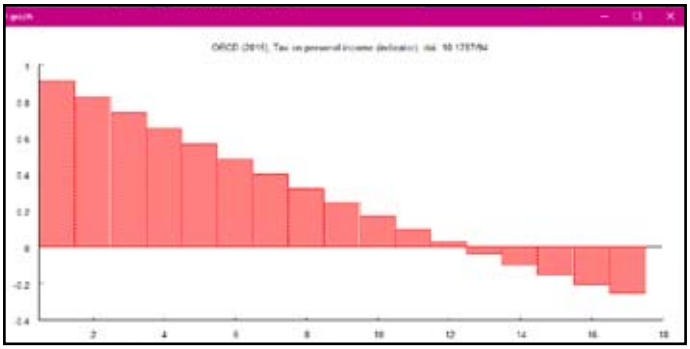

Fig. 7 : Correlogram Plot

We notice that like mirror

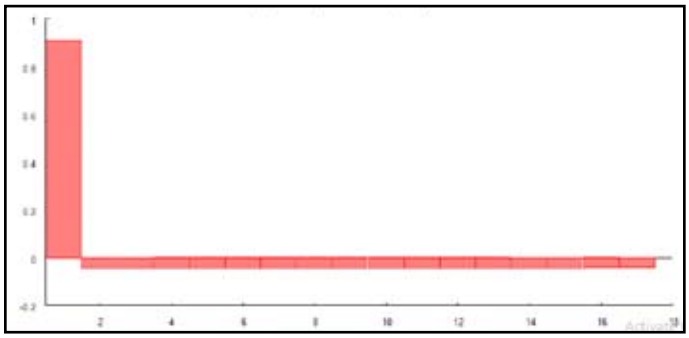

Fig. 8 : Partial correlogram

We notice that the figure is straight line.

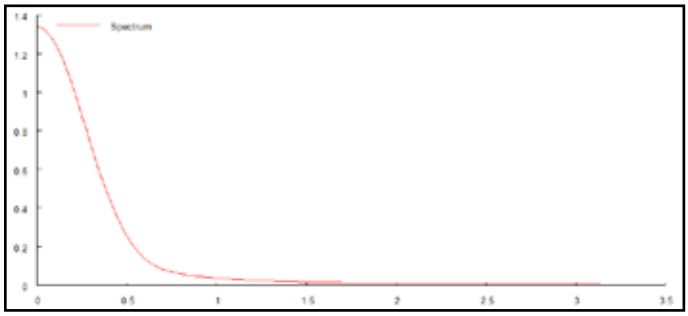


Fig. 9 : spectrum

We notice that data decrease taxes and straight.

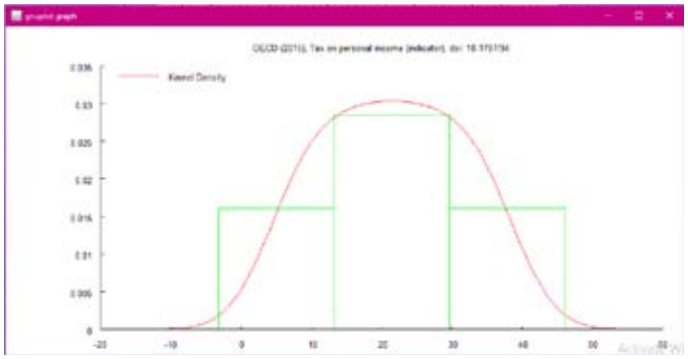


Fig.10 : Histogram /Density

The diagram shows decreased the data then increased. Linear Regression TO load Model Manger

```
**************************************************
Loading model: Defaults
Description:
  Model Run_0
  30-07-2016 at 17:49:31
  INPUT_FILE = "TAXINCOME_TOT_PC_GDP_A.xls";
  Model Settings:
  AR_ORDER = 1;
  DGTEST_LAGS = 0;
  DGTEST_SQLAGS = 0;
  IS_ARFIMA = TRUE;
  LISTING_FILE = "Run_0.tod";
  Q_TEST = LBQ;
  MA_ORDER = 1;
  METHOD = LSO;
  Q_TEST_ORDER = 0;
  SCORE_TEST = TRUE;
  Simulation Settings:
```

Fig. 11 : Producing Dynamic Equation

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Use Gaussian and use select dependent variables and select type 1

Evaluation criteria

```
*******************************************************************************
TSM4.40.16-03-16 Run 19 at 11:57:48 on 15-08-2016
Data file is
G:\Master\master\last\data(OSCD)\2003\TAXINCOME.xls
-------------------------------------------------------------------------------
Dependent Variable is OECD (2015), Tax on personal income (indicator). doi: 10.1787/94
34 observations  (1-34) used for estimation.
Estimation Method: Conditional FIML (Time Domain)
Gaussian Likelihood

Strong convergence
iteration time:  0.04
                          Estimate  Std. Err.   t Ratio  p-Value
Intercept                    21.5    1.78748    12.027        0
Error Variance^(1/2)         9.61071    0.7954    ------   ------
                    Log Likelihood = -125.882
                  Schwarz Criterion = -129.408
              Hannan-Quinn Criterion = -128.403
                  Akaike Criterion = -127.882
              Residual Sum of Squares =  3272.5
                         R-Squared =       0
                     R-Bar-Squared = -0.0312
                       Residual SD = 10.1126
                  Residual Skewness =      -0
                  Residual Kurtosis =  1.7979
                   Jarque-Bera Test =  2.0471 (0.359)
Covariance matrix from robust formula.
   ...Run completed in  0.09
```

Fig.12 : the final Result

## VI. Conclusion

This paper has investigated how tax administrations could make use of data mining and econometrics to handle inherent data problems such as missing and chronological data. Analysis of time series has been achieved by regression analysis. The operational framework is proposed. One Possible Future work direction is to try using other data types such as Panel data. Other Future work direction is trying to solve major Problems such as Trends and seasonality.

## References

[1] *Huhtanen Tiina-Liisa, Quality Manager, Egyptian Tax Administration, Helsinki Lehtinen Heli, Director, Egyptian Tax Administration, Helsinki*

[2] *Barry Anderson, TAX EXPENDITURES IN OECD COUNTRIES, Bangkok January 10-11, 2008*

[3] *L_eon Bottou, Feature engineering, COS 424 , 4/22/2010.*

[4] *Boston, OUTLIER ANALYSIS ,Kluwer Academic Publishers*

[5] *Avril Coghlan, A Little Book of R For Time Series, July 29, 2016.*

[6] *Pamela Castellon Gonzalez and Juan D. Velasquez, Characterization and detection of taxpayers with false invoices using data mining techniques, Expert Systems with Applications, no 40, pp 1427-1436, Elsevier, 2013*

[7] *Stigler, S. M., The History of Statistics. Cambridge,MA: Harvard University Press. 1986*

[8] *Kenneth Benoit_Methodology Institute London School of Economics, Linear Regression Models with Logarithmic Transformations, March 17, 2011.*

[9] *Franke R. Scattered Data Interpolation: Test of Some Methods [J]. Mathematics of Computations, 1982,33(157):181.*

[10] *Optimizing Mining Taxation for the Mineral Industry, René Artist, 2009.*

[11] *A. M. Riad, Hazem M. El-Bakry, and Gamal H. El-Adl, "A New Approach for Computing Housing Tax Rates," International Journal of Computers, vol. 6, issue 1, 2012, pp. 319-323.*

[12] *A. M. Riad, Hazem M. El-Bakry, and Gamal H. El-Adl, "A Novel Method for Determination of Housing Tax Rates," Proc. of 10th International Conference on Finance and Accounting (ICFA'10), Greece, December 29-31, 2010, pp. 9-15.*